Cleaning & Filtering in Tabular Data Analysis

Data filtering and cleaning are essential preprocessing steps in handling tabular data. These processes aim to improve data quality and, consequently, the effectiveness of data analysis or predictive models.

Data Filtering

1. **Removing Irrelevant Columns (Features)**
    1.1. Identification of Irrelevant Features:
        1.1.1. Focus on features that do not contribute to the analysis or predictive modeling. This can include redundant data, features with no variance, or data unrelated to the objective.
    1.2. Techniques for Feature Removal:
        1.2.1. Manual Removal: Based on domain knowledge.
        1.2.2. Statistical Methods:
            1.2.2.1. Numerical features:
                1.2.2.1.1. Removing Features with Low Variance
                1.2.2.1.2. Removing Features with high p-values.
            1.2.2.2. Categorical Features:
                1.2.2.2.1. Cramer's V Statistic
                1.2.2.2.2. Chi-Squared test
        1.2.3. Duplicate Data Removal
2. **Removing Irrelevant Rows (Instances)**
    2.1. Identify outliers using statistical techniques
        2.1.1. Z-score
        2.1.2. IQR
        2.1.3. DBSCAN
        2.1.4. K-means
        2.1.5. Local Outlier Factor (LOF)
        2.1.6. Isolation Forest

3. Handling Missing Values(filtering):
    3.1. Consider removing rows with excessive missing values, depending on the amount of missing data and the importance of the instances.

Data Cleaning

1. Data Validation
    1.1. Rules
        1.1.1. Range Checks: Ensure numerical values fall within reasonable and expected ranges.
        1.1.2. Format Checks: Verify that data follows expected formats (e.g., email addresses, phone numbers).
        1.1.3. List Checks: Confirm data values against a predefined list (e.g., country names, product codes).
    1.2. Data Type Checks : Confirm that each column contains the appropriate data type (e.g., numerical, string, datetime)

2. Handling Missing Data: Use strategies like imputation, deletion, or algorithms that support missing data.
    2.1. Numerical Features
        2.1.1. Dummy values
        2.1.2. One dimensional operators:
            2.1.2.1. Mean
            2.1.2.2. Median
            2.1.2.3. Max
            2.1.2.4. Min
        2.1.3. Advanced Imputation Techniques:
            2.1.3.1. K-Nearest Neighbors (KNN) imputation.
            2.1.3.2. Regression based imputation
            2.1.3.3. Multiple imputation using chained equations (MICE).
            2.1.3.4. SVD based imputation
    2.2. Categorical Features
        2.2.1. One dimensional operators
            2.2.1.1. Mode
            2.2.1.2. Dummy
        2.2.2. Advanced categorical imputation techniques
            2.2.2.1. Classification (Classification methods that can handle missing values, without the target)

2.2.2.2. K-Nearest Neighbors (KNN)

2.2.2.3. Collaborative Filtering for Imputing Missing Data

2.3. Multiple features missing values:

2.3.1.1. Multiple Imputation by Chained Equations (MICE)

2.3.1.2. Matrix Factorization for Multivariate Imputation for example Matrix Completion by Truncated Nuclear Norm Regularization

2.3.1.3. XGB/CATBOOST :The idea is to treat each column with missing data as a target variable and use the other columns as features to predict the missing values. This approach is particularly effective because these models can handle complex relationships between features.